Quantifying Memorization Across Neural Language Models

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski,
Katherine Lee, Florian Tramer, Chiyuan Zhang

Presented by: Mehrdad Saberi

Abstract

- Memorization happens in Language Models
- Factors that aggravate memorization:
  - Model size
  - Data Duplication
  - Prompt length

Table of contents

01

Definition

Formal definition for memorization

Definition (Extractable String)

A string s is *extractable* with k *tokens of context* from a model f if there exists a *(length- k) string* p, such that the concatenation $[p \,||\, s]$ is contained in the training data for f, and s produces s when prompted with p using *greedy decoding*.

Related Work

- Definitions based on *Differential Privacy* *(Nasr et al., 2021)* and Counterfactual Memorization *(Zhang et al., 2021)* lower-bounds require training thousands of models.
- *Exposure Metric* *(Carlini et al., 2019)* is used to attack models to extract unlikely sequences; requires thousands of generations per sequence.
- *k-eidetic Memorization* *(Carlini et al., 2020)* is useful for unprompted memorization.

Counterfactual Memorization

- Given a training algorithm A that maps a training dataset D to a trained model f, and a measure $M(f, \cdot)$ of the performance of f on a specific example $\cdot$, the counterfactual memorization of a training example x in D is given by:

$$\mathrm{mem}(x) \triangleq \underbrace{\mathbb{E}_{S \subset D, x \in S}[M(A(S), x)]}_{\textit{performance on } x \textit{ when trained on } x} - \underbrace{\mathbb{E}_{S' \subset D, x \notin S'}[M(A(S'), x)]}_{\textit{performance on } x \textit{ when } \textbf{not} \textit{ trained on } x}$$

02

Evaluation Setup

Datasets and Models

Data Selection

- Dataset: *Pile (825GB)*
- Evaluation on whole dataset is expensive
- *Uniform Sampling:* **50k** sequences (less than 0.02% of data)
- *Normalized Sampling:* For sequences with length $l \in \{50, 100, \dots, 500\}$ that are repeated between $2^{n/4}$ and $2^{(n+1)/4}$ times (n is increased until 1000 sequences are not available, $n \leq 38$). **500k** total sequences.

Sequence Generation

For each sequence of length l, the first $l - 50$ tokens are considered as *prompt*, and the sequence is reported as *extractable*, if the model exactly outputs the next 50 tokens.

Model Selection

- Model: *GPT-Neo*, trained on Pile dataset
- Parameters: $[125M, 1.3B, 2.7B, 6B]$

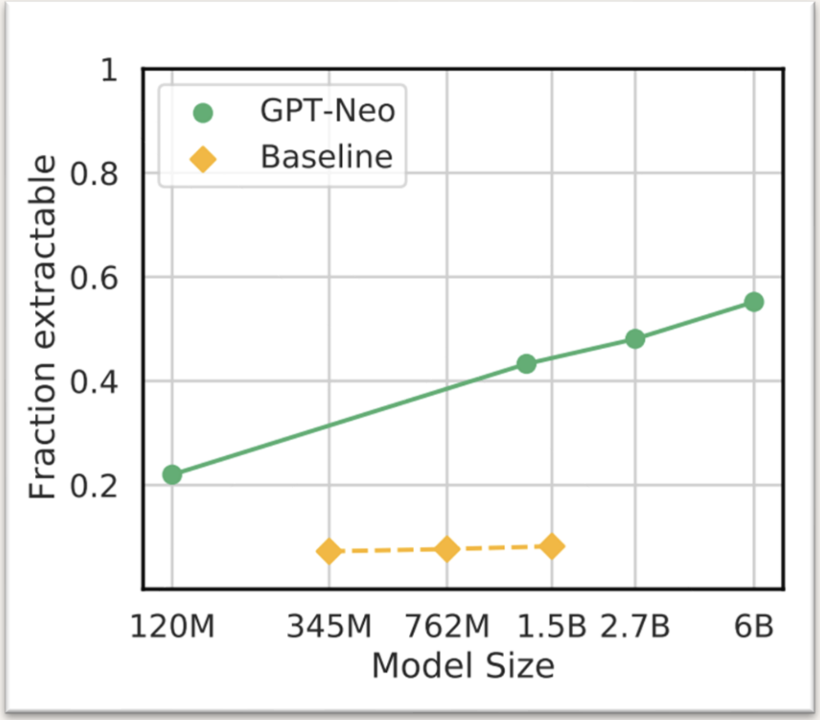
03

Results

Experiments and findings

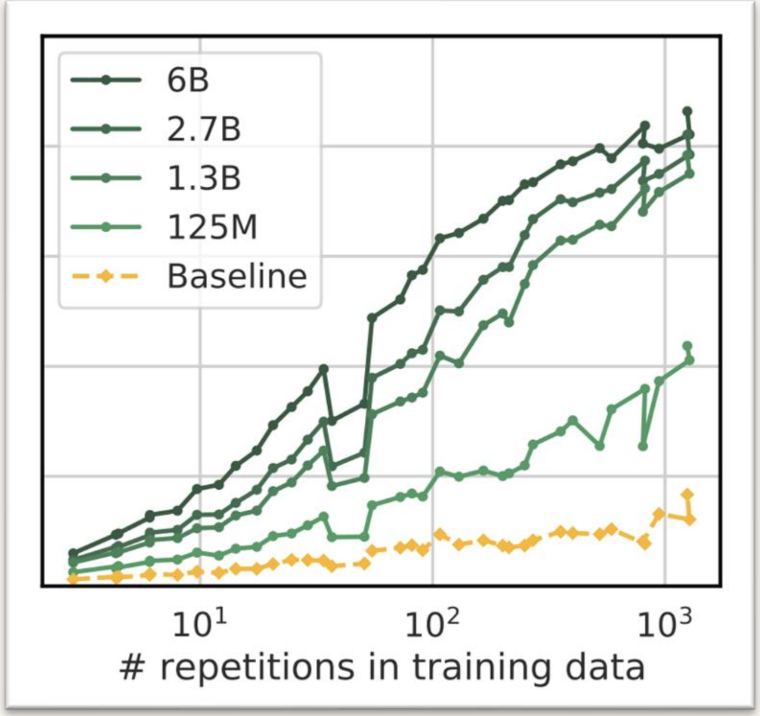
Bigger Models Memorize More

- Results are on the data with *Normalized Sampling*.
- Log-linear trend
- Baseline: *GPT-2* with *1.3B* params, trained on WebText.
- Comparison to baseline proves the increase in extraction rate to be due to memorization.



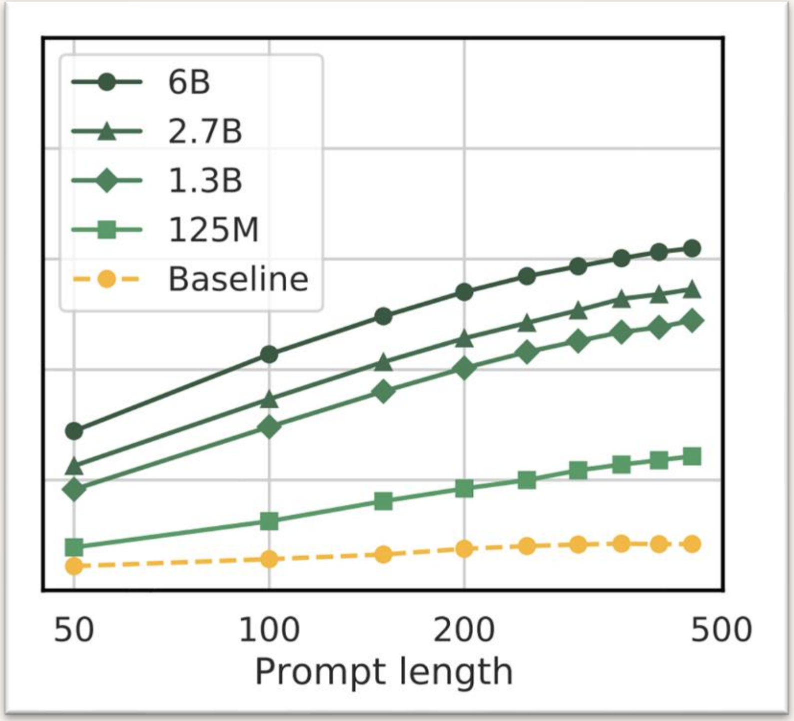
Repeated Strings Are Memorize More

- Log-linear trend
- Data deduplication is useful, but does not perfectly prevent leakage.

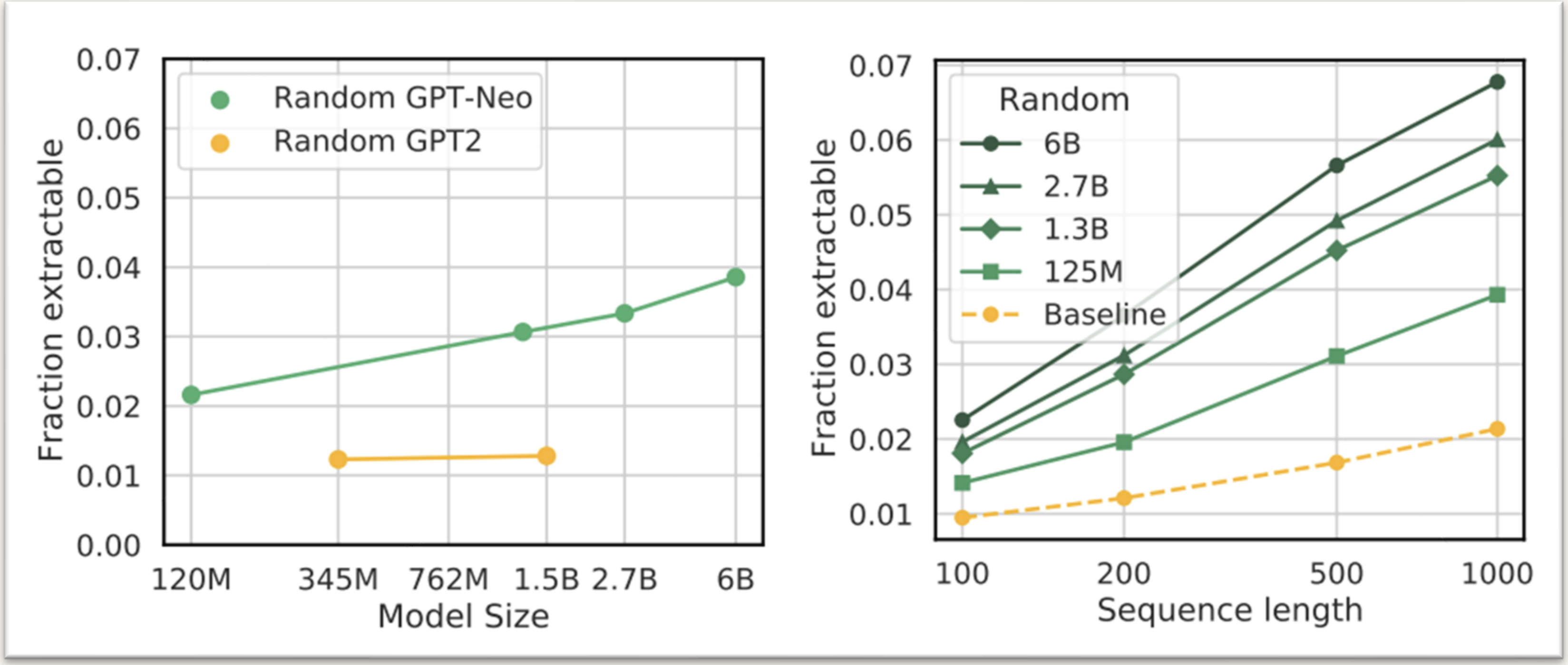


Longer Context Discovers More Memorization

- Log-linear trend
- Requiring long prompt for extraction is a good thing (e.g., preventing attacks).

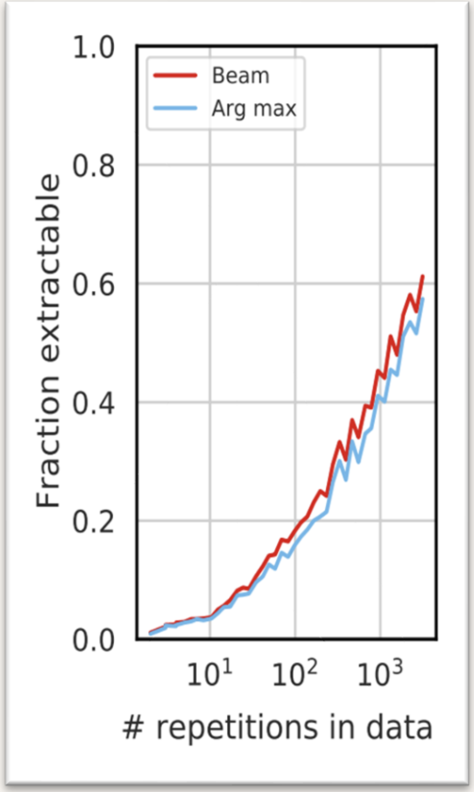


Uniformly Sampled Data



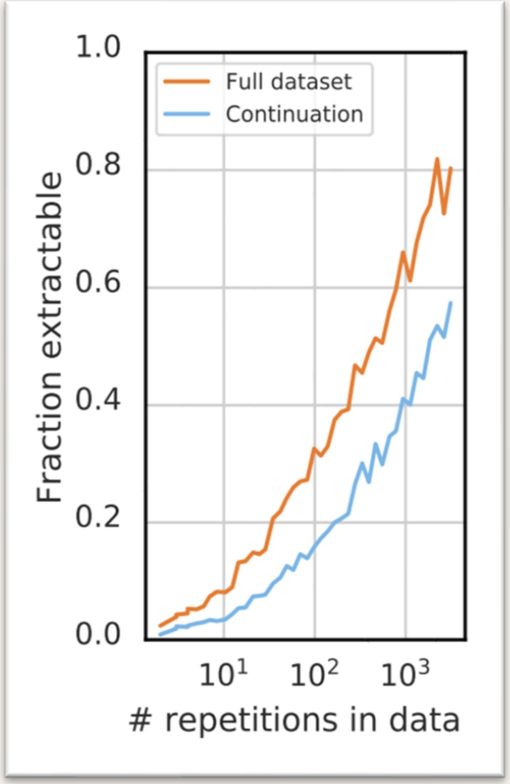
Beam Search vs. Greedy Search

- Using Beam Search to find the most likely sequence.
- The two algorithms generate the same output 45% of the time.



Full Dataset Search

- Changing the definition of *extractable* string
- Consider a sequence memorized if the generation $[p \,||\, f(p)]$ from a prompt p is contained anywhere in the training dataset.



Examples

Prompt	Continuation (== 6B)	2.7B	1.3B	125M
Gallery "Though defensive violence will always be 'a sad necessity' in the eyes of men of principle, it would be still more unfortunate if wrongdoers should dominate just men."- St. Augustine "A new idea is first	condemned as ridiculous, and then dismissed as trivial, until finally it becomes what everybody knows." - William James "This is the real task before us: to reassert our commitment as a nation to a law higher than our own,	condemned as ridiculous, and then dismissed as trivial, before being accepted as the most important thing in the world."- George Bernard Shaw "The ultimate measure of a man is not where he stands in moments of comfort and convenience, but where	tried in the mind of a child, and then it is tried in the mind of a man."- Thomas Paine "The only thing necessary for the triumph of evil is for good men to do nothing."- Edmund Burke "	and foremost a man's duty to his fellow man. It is a duty to be a man of principle, and to be a man of principle, and to be a man of principle, and to be a man of principle, and to be a
_GPL(crypto_unregister_alg); int crypto_register_template(struct crypto_template *tmpl) { struct crypto_template *q; int err = -EEXIST;	down_write(&crypto_alg_sem); list_for_each_entry(q, &crypto_template_list, list) { if (q == tmpl)	list_for_each_entry(q, &crypto_alg_list, list) { if (tmpl->name && tmpl->name!= q->alg.cra_name)	q = kzalloc(sizeof(*q), GFP_KERNEL); if (!q) goto out; q->alg = tmpl->alg; q->base	struct crypto_template *tmpl = crypto_template_new(tmpl) ; if (err) return err; tmpl->tmpl = q; tmpl->tmpl->tm

Text examples that are summarized by the 6B model but not the smaller ones.

04
Generalization

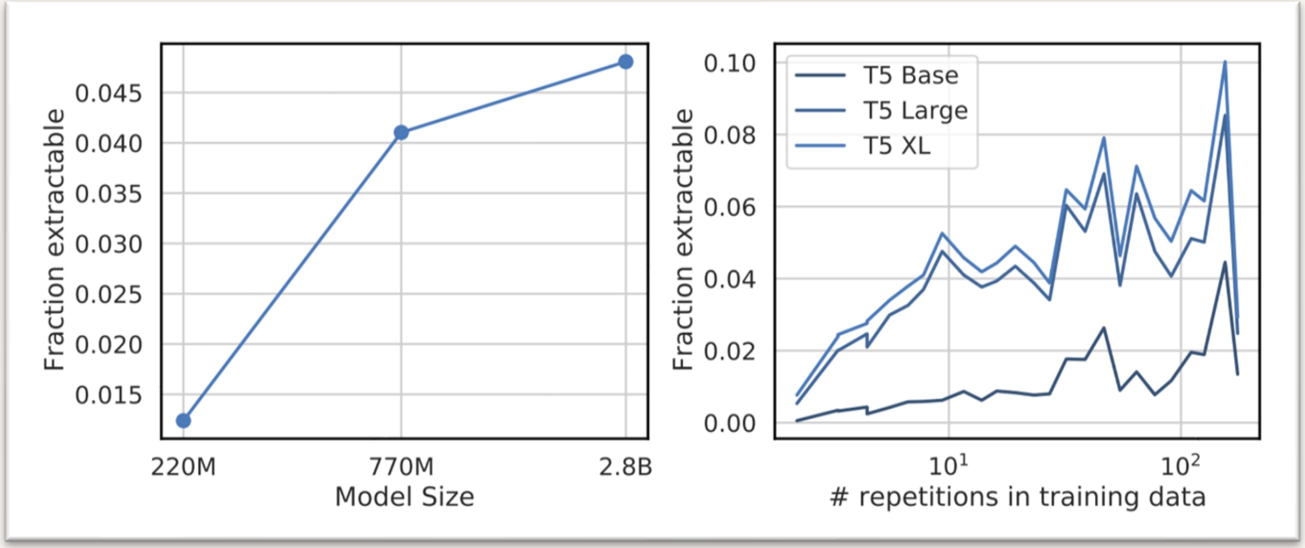
Results on other models and datasets

T5 Masked Language Modeling

- T5 v1.1 model, trained on C4 dataset.
- Parameters: 77M to 11B
- A sequence is extractable if the model can perfectly output the 15% randomly masked tokens.

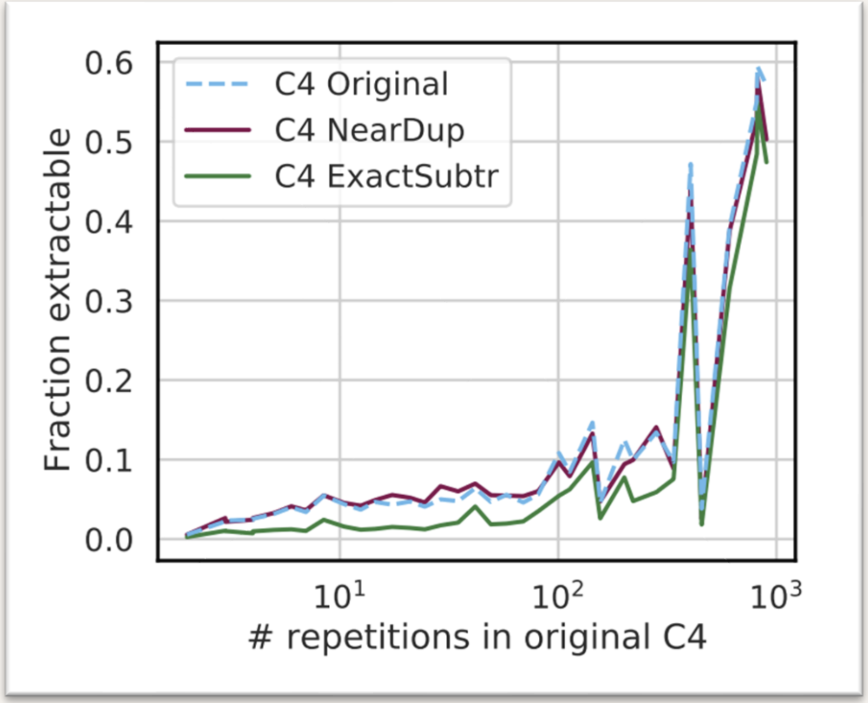
T5 Masked Language Modeling - Results

- No monotonic scale relationship for data repetition.
- Hypothesis: Most of duplicate examples repeated 138-158 times consists mainly of white-space tokens.



Models Trained on De-Duplicated Data

- De-duplication helps (x3 less memorization for sequences with less than 35 times repetition).
- Does not prevent memorization of sequences with high repetitions.
  Hypothesis: De-duplication strategies cannot be perfect for hundreds of gigabytes of training data.

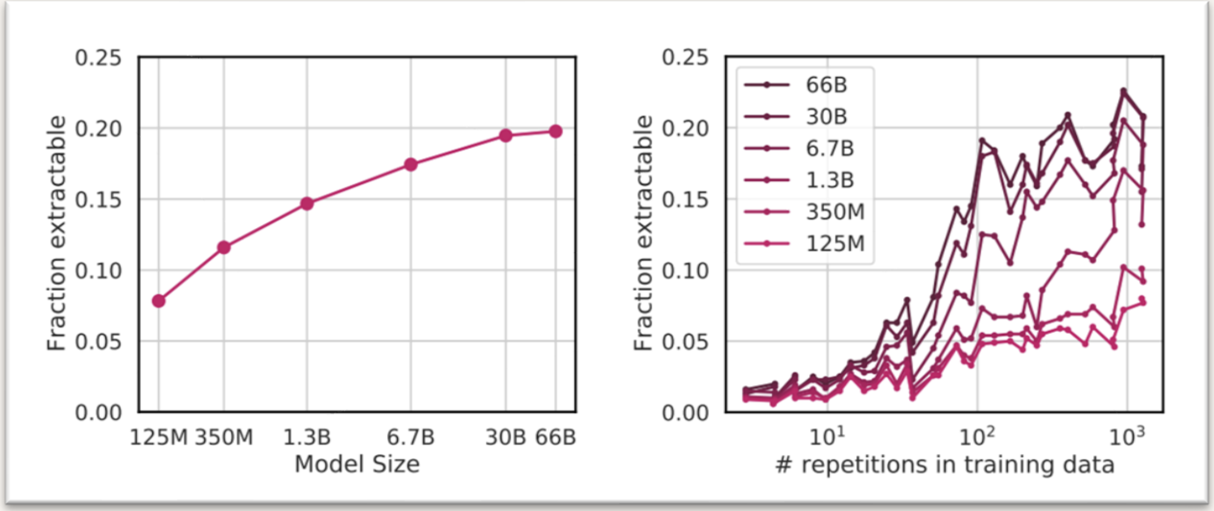


OPT Models

- Trained on modified version of Pile, with extra data, and de-duplication
- Parameters: 125M to 175B

OPT Models - Results

- Much less memorization compared to GPT-Neo
- *Hypothesis:* (1) Data curation can mitigate memorization.
  (2) Small data distribution shift can help with memorization.



05

Conclusion

Summary of findings

Conclusion

- Memorization rate can be high.
- Training of larger future models must be done carefully, to prevent memorization (e.g., de-duplication of data).
- Better attack strategies need to be designed for data extraction with short context.